STA237 Notes

Probability without Proofs

https://github.com/ICPRplshelp

Last updated December 29, 2022

1 Probabilities

Unions and intersections

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$$

Unions for nonempty intersections

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Combinations and permutations

$$P_r^n = \frac{n!}{(n-r)!}$$
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Givens

$$P(A|C) = \frac{P(A \cap C)}{P(C)}, \text{ 0 if } P(C) = 0$$

The multiplication rule

$$P(A \cap C) = P(A|C)\,P(C)$$

Bayes

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

The law of total probability

$$P(A) = \sum_{i=1}^{k} P(A|B_i)\,P(B_i)$$

2 Expected Values and Variance

$P(X = 8)$ is the probability X happens to be 8.

Expected value:

$$E(X) = \sum_{\forall x} xP(x) = \mu$$

Where $\sum_x xP(x)$ tries to produce a weighted average of all possible outcomes. No divisions because that was already accounted for in $P(x)$.

Variance:

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right] = \sum_x (x - E(X))^2 P(x)$$

Expected values are linear:

$$E(aX+b) = aE(X)+b$$

Variance: taking anything out of V that multiplies the random variable requires you to square it. $V(bX+a) = b^2V(x),\ \forall a,\ b \in \mathbb{R}$.

3 Bernoulli Trials and Binominal Distributions

A trial that can fail or succeed is called a Bernoulli trial. Then, X can take 0 or 1, and

$$P(\text{Success}) = P = P(X=1)$$
$$P(\text{Fail}) = 1 - P = P(X=0)$$

The expected value of a Bernoulli distribution is $E(X) = P$ and its variance is $V(X) = 1 - P$. When a variable follows a Bernoulli distribution, $X \sim Ber(p)$.

4 Binominal distribution

> Use this if you want to accuse someone of cheating whilst assuming that the "stopping criterion" never existed.

X follows this distribution if given $n \in \mathbb{N}^{\geq 1}$ and $p \in [0, 1]$, its PMF is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where n is the number of trials and x is the number of successes.

Use the binominal distribution table to figure out what $P(X \leq k)$ is for n trials.

You may assume independence if we take a sample size that will always be less than 10% of the entire population – then if we sampled them without replacement we can treat as if that we sampled them with replacement.

4.1 Expected values and variance

$$E(X) = np$$
$$V(X) = np(1-p)$$

(Magnified by number of trials)

5 Geometric Distribution

Use it when you want to calculate the number of required trials **each being independent of each other** until first success, and p is the probability of success each time we try, then if $y \in \mathbb{N}^{\geq 1}$ is the number of times we rerolled the dice **(counting from 1)**, the probability that our first success will be on the yth roll will be:

$$P(Y = y) = (1 - p)^{y-1}$$

To safely assume independence on something that clearly is dependent, if the quantity we choose from (y) is less than 10% of the quantity in the entire box then we can assume independence.

5.1 Example situation

A question that would require this to solve it might look like:

> If I start sampling something from this box **with replacement** until I get something with a particular trait, what is the probability that I will stop ONLY on the 5$^{\text{th}}$ sample?

Then $y = 5$, and p is the probability that I'll observe that trait regardless.

5.2 Expected values and variance

The expected value and variance for Y if it follows a geometric distribution is:

$$\mu = E(Y) = \frac{1}{p}$$
$$\sigma^2 = V(Y) = \frac{1-p}{p^2}$$

6 Hypergeometric Distributions

For times where you'll have to use the binominal distribution but cannot assume independence, i.e., **you are sampling without replacement and the sample size is small.**

In our context, if we define our variables like these:

- N: entire sample size (size of the box)
- n: number of times we'll sample from the box
- r: number of successes in the entire sample size (entire box)
- x: number of successes in our sample (from what we've picked out)

Then:

$$P(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Note that due to the pigeonhole principle and that we can't take more than what's in the box, the number of successes has the following constraints:

- Can't be lower than $\max\left(0,\ n-(N-r)\right) = \max\left(0,\ n - \text{failures in entire box}\right)$
- Can't be higher than $\min(r,\ n) = \min\left(\text{no. successes in box},\ \text{size of box}\right)$

As a domain, we can describe this as:

$$x \in \left[\max\left(0,\ n-(N-r)\right), \min(r,\ n)\right]$$

6.1 Expected value and variance

$$E(X) = n \cdot \frac{r}{N}$$
$$V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{r}{N} \cdot \left(1 - \frac{r}{N}\right)$$

7 The Poisson distribution

> If an event happens independently and randomly over time and the mean rate of occurrence is constant over time, then the number of occurrences in a fixed amount of time will follow the Poisson distribution. As a random variable, it would be X. (Source: HERE)

If X follows a Poisson distribution, then EVs and variance match:

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

And

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x \in [0,\ \infty)$$

8 Continuous random variables

- The continuous **density** function is $f(y)$.
- The continuous **distribution** function is $F(y) = \int_{-\infty}^{y} f(t)dt$.

A rule for all continuous distribution functions: $F(\infty) = 1$ and $F(-\infty) = 0$. Unless the event is guaranteed or is zero for some weird reason.

$$P(a \le Y \le b) = \int_a^b f(y)dy = F(b) - F(a)$$

Expected value and variance

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f(y)dy$$
$$E(g(Y)) = \int_{-\infty}^{\infty} g(y) \cdot f(y)dy$$
$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} y^2 \cdot f(y)dy$$

8.1 Uniform distribution

A continuous distribution is uniform if it is constant in an interval $[a, b]$, 0 otherwise.

$$f(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

8.1.1 Expected value and variance

$$\mu = E(X) = \frac{b-a}{2}$$
$$\sigma^2 = V(X) = \frac{(a-b)^2}{12}$$

8.2 Normal distribution

The probability distribution function for anything that follows a normal distribution is:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

If we say that something follows a normal distribution, then we can say

$$Y \sim N(\mu, \sigma)$$

The expected value and variance is exactly where you see it in the normal distribution formula.

8.2.1 The standard normal (Zs)

$$Z \sim N(0, 1)$$
$$\Rightarrow P(Z = z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Without a calculator, as long as you have the standard normal distribution table, this formula holds, where $X \sim N(\mu, \sigma)$:

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

Cool trick:

$$\begin{aligned} P\left(-z_0 < Z < z_0\right) &= a \\ \Leftrightarrow P\left(Z < z_0\right) &= a + \frac{1-a}{2} \end{aligned}$$

9 Binominal to normal Approximation

One of the applications of the central limit theorem. I have no idea why, but as $X \sim \text{Ber}(n,\ p)$, and I'm doing a ton of independent re-rolls, it fits the situation.

Use this if, in a binominal distribution, **both** the expected number of successes np and the expected number of failures $n(1-p)$ exceed 10. Then, use the normal distribution which retains the same expected value and standard deviation (remember that standard deviation is used for the normal distribution).

Also, the **continuity correction must always be used.** That is:

$$P(a \leq X \leq b) = P\left(a - \frac{1}{2} \leq Y \leq b + \frac{1}{2}\right) = \int_{a-\frac{1}{2}}^{b+\frac{1}{2}} f(y)dy$$

10 Gamma Distribution

Gamma density function

$$f(y) = \frac{y^{\alpha-1}e^{-\frac{y}{b}}}{\beta^\alpha \Gamma(\alpha)}$$

Where $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for any $\alpha > 1$. This means that

$$\Gamma(n) = (n-1)!$$

And here's the form of the gamma function in the form of an integral, which will be useful:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

Note that y is just a variable used for integration – we could realistically take any integral that looks like this and substitute it with the gamma function.

Be sure to be able to pattern-recognize any integral that looks like I'm multiplying a polynomial by a flipped exponential!

10.1 Derivation of the expected value of the Gamma distribution

The expected value of anything that follows the Gamma distribution is:

$$\mu = E(Y) = \int_0^\infty \frac{y \cdot y^{\alpha-1}e^{-\frac{y}{\beta}}}{\beta^\alpha\Gamma(\alpha)}dy$$

Let $\zeta = \frac{y}{\beta}$. Then, $y = \beta\zeta$ and $dy = d\beta\zeta$.

Derivation.

$$
\begin{aligned}
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} y \cdot y^{\alpha-1} e^{-\frac{y}{\beta}} dy && \text{move the denominator out} \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} y^{\alpha} e^{-\frac{y}{\beta}} dy && \text{exponent rule for } y \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} (\beta\zeta)^{\alpha} e^{-\zeta} d\beta\zeta && \text{substitute zeta} \\
&= \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} \zeta^{\alpha} e^{-\zeta} d\zeta && \text{cancel out } \beta^{\alpha} \text{ and move out right } \beta \\
&= \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} \zeta^{(\alpha+1)-1} e^{-\zeta} d\beta\zeta && \text{make it lool like } \Gamma(\alpha+1) \\
&= \frac{\beta}{\Gamma(\alpha)} \Gamma(\alpha+1) && \text{substitute } \Gamma(\alpha+1) \\
&= \frac{\beta}{\Gamma(\alpha)} \alpha \cdot \Gamma(\alpha) && \text{use the property of } \Gamma(\alpha+1) \\
&= \beta \cdot \alpha && \text{cancel out } \Gamma(\alpha)
\end{aligned}
$$

We can conclude that $\mu = E(Y) = \beta\alpha$.

10.2 Derivation of the variance of the Gamma distribution

Our declarations of $\zeta = \frac{y}{\beta}$, $y = \beta\zeta$, and $dy = d\beta\zeta$ remain. To find $V(Y)$, we first need to find $E\left(Y^2\right)$. Which is:

Derivation.

$$
\begin{aligned}
E\left(Y^2\right) &= \int_0^\infty \frac{y^2 \cdot y^{\alpha-1} e^{-\frac{y}{\beta}}}{\beta^\alpha \Gamma(\alpha)} dy && \text{setup} \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty y^{\alpha+1} e^{-\frac{y}{\beta}} dy && (y^2 \cdot y^{\alpha-1} = y^{\alpha+1}) \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty (\beta\zeta)^{\alpha+1} e^{-\zeta} d\beta\zeta && \text{sub } y = \beta\zeta \\
&= \frac{\beta^2}{\Gamma(\alpha)} \int_0^\infty \zeta^{\alpha+1} e^{-\zeta} d\zeta && \text{cancel } \beta \text{ out} \\
&= \frac{\beta^2}{\Gamma(\alpha)} \int_0^\infty \zeta^{(\alpha+2)-1} e^{-\zeta} d\zeta && \text{make it look like } \Gamma(\alpha) \\
&= \frac{\beta^2}{\Gamma(\alpha)} \Gamma(a+2) && \text{sub } \Gamma(\alpha+2) \\
&= \frac{\beta^2}{\Gamma(\alpha)} (\alpha+1) \cdot \alpha \cdot \Gamma(\alpha) && \text{use the property of } \Gamma(\alpha+2) \\
&= \alpha^2\beta^2 + \alpha\beta^2 && \text{cancel out } \Gamma(\alpha) \\
&= E\left(Y^2\right) && \text{conclusion}
\end{aligned}
$$

There we go. Because $V(Y) = E\left(Y^2\right) - (E(Y))^2$, we can substitute:

$$
\begin{aligned}
V(Y) &= \alpha^2\beta^2 + \alpha\beta^2 - (\beta\alpha)^2 \\
&= \alpha\beta^2
\end{aligned}
$$

11 Exponential Distribution

The exponential density function is, with $\beta > 0$:

$$f(y) = \frac{1}{\beta} e^{-\frac{y}{\beta}},\ y \geq 0,\ 0 \text{ otherwise}$$

11.1 Expected value of the exponential distribution

The expected value $E(Y)$ is:

$$\begin{aligned}
E(Y) &= \int_0^\infty y \cdot \frac{1}{\beta} e^{-\frac{y}{\beta}} dy && \text{setup} \\
&= \frac{1}{\beta} \int_0^\infty y \cdot e^{-\frac{y}{\beta}} dy && \text{move } \frac{1}{\beta} \text{ out} \\
&= \frac{1}{\beta} \left(\left[\frac{ye^{-y\beta}}{\frac{1}{\beta}} \right]_\infty^0 - \int_0^\infty \frac{e^{-\frac{y}{\beta}}}{\frac{1}{\beta}} dy \right) && \text{integrate by parts} \\
&= \left[ye^{\frac{y}{\beta}} \right]_\infty^0 - \int_0^\infty e^{-\frac{y}{\beta}} dy && \text{cancel } \frac{1}{\beta} \text{ out} \\
&= 0 - 0 + \left[\frac{e^{-\frac{y}{\beta}}}{\frac{1}{\beta}} \right]_\infty^0 && \text{solve integral again} \\
&= \beta(1-0) && \text{using l'hopital's rule} \\
&= \beta = E(Y) && \text{conclusion}
\end{aligned}$$

You may need to use the “big theorem.”

11.2 Variance of the exponential distribution

To calculate $V(Y) = \sigma^2$, we first need to figure out $E\left(Y^2\right)$:

$$
\begin{aligned}
E\left(Y^2\right) &= \int_0^\infty y^2 \cdot \frac{1}{\beta} e^{-\frac{y}{\beta}} dy && \text{setup} \\
&= \frac{1}{\beta} \int_0^\infty y^2 e^{-\frac{y}{\beta}} dy && \text{take } \frac{1}{\beta} \text{ out} \\
&= \frac{1}{\beta} \left(\left[\frac{y^2 e^{-\frac{y}{\beta}}}{\frac{1}{\beta}} \right]_\infty^0 + \int_0^\infty \frac{e^{-\frac{y}{\beta}}}{\frac{1}{\beta}} \cdot 2y \, dy \right) && \text{integrate by parts} \\
&= \left[y^2 e^{-\frac{y}{\beta}} \right]_\infty^0 + 2\beta \int_0^\infty \frac{e^{-\frac{y}{\beta}} \cdot 2y}{\beta} dy && \text{cancel } \frac{1}{\beta} \text{ out} \\
&= 0 + 2\beta \int_0^\infty \frac{y e^{-\frac{y}{\beta}}}{\beta} dy && \text{solve the left side} \\
&= 2\beta E(Y) && \text{recognize something?} \\
&= 2\beta^2 && \text{expand the EV}
\end{aligned}
$$

$$2\beta^2 - \beta^2 = \beta^2$$

12 Simulation

Given a random variable's **distribution** function $F(x)$, to run a simulation:

1. Invert it (get $F^{-1}(x)$)
2. Your random number generator is F^{-1} (random()), where `random()` generates a random floating-point number between 0 and 1.

More specifically, if U is a uniform $(0,\ 1)$ random variable, then $Y = F^{-1}(U)$ shares the same distribution function as F.

13 Joint Probability Functions

The joint probability density function is just a $\mathbb{R}^2 \to \mathbb{R}$ map. y_1 may be one factor, y_2 may be another. Its domain may be limited, and the entire volume of it must sum up to 1. Any point (x, y) that is nonzero in the joint probability density function is called the active region.

As a rule of thumb, dy_2 goes in the inner integral, and dy_1 goes in the outer integral. Although the order doesn't matter, it feels cleaner to do this. Hence, you look **up/down** first before looking **left/right** afterwards.

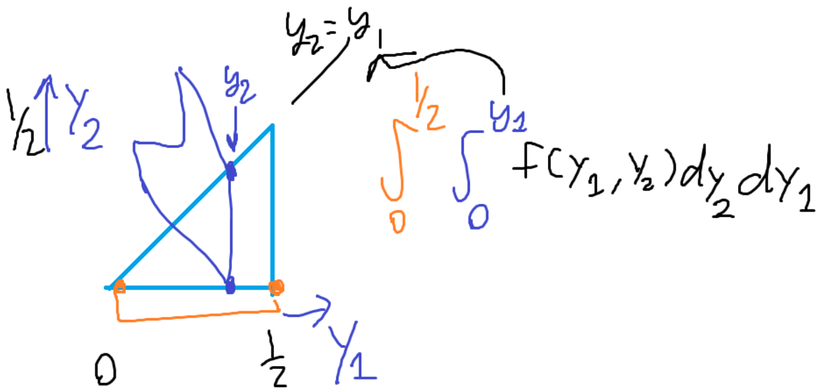

Figure 1: How double integrating looks like

14 Marginal Distribution for Bivariate

> Marginal probability is **the probability of an event irrespective of the outcome of another variable.** - SOURCE.

Example: I know, for each person the time they studied and the percent of questions they got correct. Now, how many of them got between 20-40% correct, if I don't know how long they studied for?

For discrete variables, it is:

$$p_1(y_1) = \sum_{\forall y_2} p(y_1, y_2)$$

y_1 is fixed and is an argument of $p_2(y_1)$

$$p_2(y_2) = \sum_{\forall y_1} p(y_1, y_2)$$

y_2 is fixed and is an argument of $p_2(y_2)$. Since our outcome isn't affected by Y_1, we sum up what could've happened for any Y_1.

For continuous (the formulas are similar), it is:

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\, dy_2$$

To find a vertical slice, and

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2)\, dy_1$$

To find a horizontal slice.

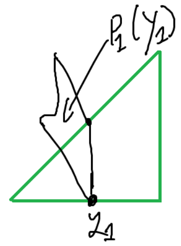

Figure 2: Marginal probability

15 Marginal and Conditional Probability Distributions

15.1 Discrete

If Y_1 and Y_2 are jointly discrete with the joint probability function $p(y_1, y_2)$ and marginal probability function $p_1(y_1)$ and $p_2(y_2)$:

$$
\begin{aligned}
p\left(y_1|y_2\right) = p\left(Y_1 = y_1|Y_2 = y_2\right) &= \frac{P\left(Y_1 = y_1 \wedge Y_2 = y_2\right)}{P\left(Y_2 = y_2\right)} \\
&= \frac{p_1\left(y_1, y_2\right)}{p_2\left(y_2\right)}
\end{aligned}
$$

Given that $p_2\left(y_2\right) > 0$.

15.2 Continuous

If Y_1 and Y_2 are jointly continuous, with joint density $f\left(y_1, y_2\right)$ and marginal densities $f_1\left(y_1\right)$ and $f_2\left(y_2\right)$, for any y_2 such that $f_2\left(y_2\right) > 0$, the conditional density of Y_1 given $Y_2 = y_2$ is:

$$
f\left(y_1|y_2\right) = \frac{f\left(y_1, y_2\right)}{f_2\left(y_2\right)}
$$

15.3 Imprecise

These rules don't apply to questions that ask for

$$
P\left(Y_1 < c \mid Y_2 \leq b\right)
$$

You'll have to consider the idea of conditional probability:

$$
\frac{P\left(Y_1 < c \text{ and } Y_2 \leq b\right)}{P\left(Y_2 \leq b\right)}
$$

You may need visual aids to figure out $P\left(Y_1 < c \text{ and } Y_2 \leq b\right)$. Key word: <u>and</u> means intersection.

16 Univariate Transformations

When given a single variable density function $f_Y(y)$ and you want to find the probability density function $U = h(Y)$, here are the steps:

Firstly, find $h^{-1}(u) = y$ by inverting h.

Then, use this formula:

$$f_U(u) = f_Y\left(h^{-1}(u)\right)\left|\frac{dh^{-1}}{du}\right|$$

This is literally applying the chain rule. Nothing special going on here.

Beware: you do need to find bounds, which can always be found using this method:

$$\begin{gathered} a \leq y \leq b \\ \Rightarrow h^{-1}(a) \leq u \leq h^{-1}(b) \end{gathered}$$

16.1 Without The Transformation Method

If you're asked to do a question like this but are not allowed to use the method of transformations, start with

$$\begin{aligned} P(U < u) &= P\left(h^{-1}(U) < h^{-1}(u)\right) \\ &= P\left(Y < h^{-1}(u)\right) = F\left(Y < h^{-1}(u)\right) \\ &= F\left(h^{-1}(u)\right) \end{aligned}$$

And differentiate that. You'll get:

$$\frac{d}{du}F\left(h^{-1}(u)\right) = f\left(h^{-1}(u)\right) \cdot \frac{dh^{-1}(u)}{du}$$

17 Multivariate Transformations

When given a multivariate density function in the form: $f_{Y_1, Y_2}(y_1, y_2) =$ something and I'm asked to find $f_U(u)$, there are some steps I have to do. U is a random variable which is a function of Y_1 and Y_2, maybe $h(Y_1, Y_2)$.

1. Declare U_1 and U_2 such that $U = U_1$ and $U_2 =$ some function of Y_1 and Y_2 (and the function must contain either that does not cancel out).
2. Replace the random variable version with the non-random variable version. For this case:
   a. $u_1 = h(y_1, y_2)$ and $u_2 = h(y_1, y_2)$.
3. Invert these functions. By convention, h_1^{-1} links to y_1 and h_2^{-1} links to y_2.
   a. $y_1 = h_1^{-1}(y_1, y_2)$ and $y_2 = h_2^{-1}(y_1, y_2)$.
4. Then we have this:
   a. $f_{U_1, U_2}(u_1, u_2)$
   $= f_{Y_1, Y_2}\left(h_1^{-1}(y_1, y_2), h_2^{-1}(y_1, y_2)\right) ||J||$
   b. Where $|J| = \begin{vmatrix} \frac{\partial h_1^{-1}}{\partial u_1} & \frac{\partial h_1^{-1}}{\partial u_2} \\ \frac{\partial h_2^{-1}}{\partial u_1} & \frac{\partial h_2^{-1}}{\partial u_2} \end{vmatrix}$.
   c. The double absolute value is not a typo. You'll need the absolute value of the determinant of the Jacobian matrix.
5. Compute $f_{U_1}(u_1)$. You know how to do this. You've seen it before (use marginal probability). Then, conclude with $f_U = f_{U_1}$.

Note that u_1 and u_2 are bounded. Namely:

- If $a \leq y_1 \leq y_2 \leq b$:
  - Then, $h_1(a) < u_1$ and $u_1 < u_2$ and $u_2 < h_2(b)$
- Similar arguments apply if $a \leq y_1 \leq c$ and $b \leq y_2 \leq d$.

17.1 Without The Transformation Method

If you're given the density function $f_{Y_1,Y_2}(y_1, y_2)$ and the transformation $U = h(Y_1, Y_2)$, and you're not allowed to use the method of transformations, here's what you need to do:

For $P(U < u)$ (we fix an arbitrary u):

1. Find the subset of $\mathbb{R}^2$ such that $h(Y_1, Y_2) < u$.
2. Find the active region of f_{Y_1,Y_2} (also a subset of $\mathbb{R}^2$).
3. The active region of what you want to integrate f_{Y_1,Y_2} is the **intersection** of the regions found in step 1 and step 2. You may need to apply some clever tactics when finding the required probability, like using complements or splitting areas up.

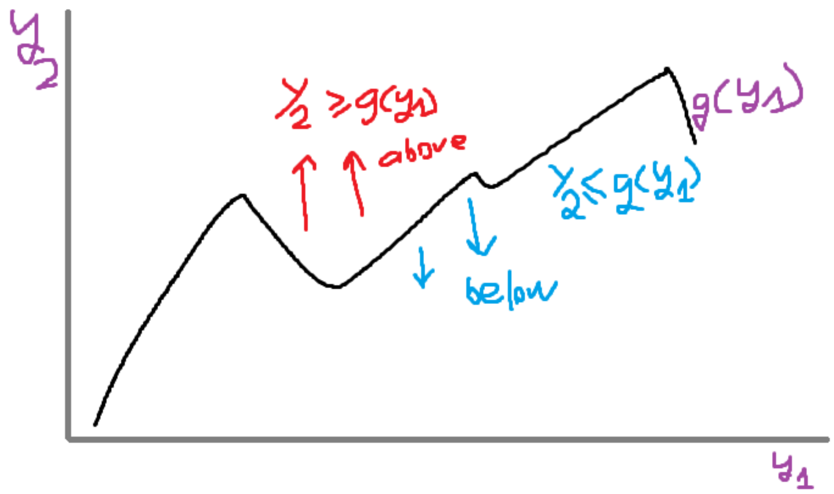

Figure 3: Visualizing inequalities

A subset of $\mathbb{R}^2$ just means area on the graph with y_1 as the x-axis and y_2 as the y-axis, for the context of this document.

18 Covariance

If Y_1 and Y_2 are random variables with means μ_1 and μ_2 respectively, the covariance of Y_1 and Y_2 is:

$$\begin{aligned} Cov\left(Y_1,\, Y_2\right) &= E\left(\left(Y_1-\mu_1\right)\left(Y_2-\mu_2\right)\right) \\ &= E\left(Y_1Y_2\right) - E\left(Y_1\right)E\left(Y_2\right) \end{aligned}$$

Plug and chug away. Note that the expected value is treated as function, so $E\left(Y_1Y_2\right)$ would work by multiplying the integrals together:

$$\iint_{\mathbb{R}_2} y_1y_2 f\left(y_1,\, y_2\right) dy_2 dy_1$$

Correlation and independence: independence $\Rightarrow$ not correlated. When the correlation is 0, there is no correlation. The correlation coefficient ρ is:

$$\rho = \frac{Cov(X,\, Y)}{\sqrt{V(X)V(Y)}}$$

19 Sampling Distribution

I use capital letters to denote a population, and lowercase letters to denote a sample.

19.1 Dealing With A Population

If I had access to an entire population with N people, and I want to observe a property from them, my observation might look like

$$Y_1,\, Y_2,\, \ldots,\, Y_N$$

The theoretical mean and the theoretical variance are "there" and are denoted by μ and σ^2. Too bad most of the time, you can't measure them especially considering that time passes.

- The theoretical mean is $\frac{1}{N}\sum_{i=1}^{N} Y_i$
- The theoretical variance is $\frac{1}{N}\sum_{i=1}^{N}\left(Y_1-\mu^2\right)$. Feel free to use properties with $\sum$s to simplify this.

19.2 Dealing With A Sample

A **random** sample (that is large enough…?) is sufficient to make a conclusion about the whole problem. Your sample **MUST** be random, otherwise sampling bias could occur (practically, this is impossible).

If our sample size is of size n, we might denote our sample findings as:

$$y_1,\, y_2,\, \dots,\, y_n$$

This is our collected random sample. Then:

- $\overline{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean
- $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}^2\right)$ is the sample variance.
  - We divide by $\frac{1}{n-1}$ because apparently, this makes $E\left(s^2\right)=\sigma^2$. Please stop asking why.

We call these the statistic.

The connection between $\mu,\ \sigma^2$ and $n,\ \overline{y},\ s^2$ is called the **sampling distribution.**

19.3 The Mean of the mean and the Standard Deviation of the Mean

I'd rather call the mean of the mean the expected value of the mean.

$$\mu_{\overline{y}} = \mu$$
$$\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}} \text{ where } n \text{ is the sample size}$$

20 The Max and Min of Multiple Random Variables / Order Statistic

If I'm running many random dice rolls, what is the density/distribution of the max of them?

If I have a sequence of random variables, each independent and sharing the same probability distribution, and I sort them such that:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(k)}$$

Then, the density function of $Y_{(k)}$ is given by:

$$g_{(k)}(y) = \frac{n!}{(k-1)!(n-k)!} (F(y))^{k-1} (1 - F(y))^{n-k} f(y)$$

21 The Central Limit Theorem

If I have a large **(30 or over)** number of independent variables with the exact same distribution $X_1, \ldots, X_n$, then $\overline{X}_n$ approximately has a normal distribution.

Okay. Here's the actual theorem:

Theorem (Central Limit Theorem). Let $n \geq 30$. Let $y_1, y_2, \ldots, y_n$ be independent identically distributed random variables such that $\forall i \in [1, n]\ E(y_i) = \mu$ and $V(y_i) = \sigma^2 < \infty$. If we define:

$$Z_n = \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{y} - \mu_{\overline{y}}}{\sigma_{\overline{y}}}$$

Then Z_n follows the standard normal distribution $Z_n \sim N(0,\ 1)$.

If σ is unknown, then replace σ with s, where s is the standard deviation:

$$Z_n = \frac{\overline{y}-\mu}{\frac{s}{\sqrt{n}}} \sim N(0,\ 1)$$

If the sample size is less than 30, then ONLY if $Y \sim N(\mu,\ \sigma)$, then $\overline{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

21.1 Sample Proportions

If you see $\widehat{p}$, that is the proportion of successes. This only works if the random variable in concern is Bernoulli. Then:

$$\begin{aligned} E\left(\widehat{p}\right) &= \mu = p \\ V\left(\widehat{p}\right) &= \frac{\sigma^2}{n} = \frac{p(1-p)}{n} \\ \sigma_{\widehat{p}} &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

Where p is the probability per trial. It's similar to dealing with a sample mean, but with a different way to get the variance and standard deviation.

If $np \geq 10$ and $n(1-p) \geq 10$, we can claim that

$$\widehat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

No need for continuity correction. Use the normal distribution.

Look for key words signaling that a random process is Bernoulli.

22 T-Distribution

Typically used to calculate the probability that the theoretical mean (used as part of the null hypothesis) is in some bounds (typically a multiple of the standard deviation) for the purposes of getting a p-value. Hence, **the probability that the distance from the sample mean and the theoretical mean is below some value.**

If sample size is small, we don't know the theoretical SD, **and it follows a normal distribution**, then:

$$T = \frac{(\bar{y} - \mu)}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Where $n-1$ is the degrees of freedom. Where n is the sample size.

When solving any problem involving the t-distribution, start off with your usual $P(|\bar{y} - \mu| < \cdots)$ (the $\cdots$ likely must be a multiple of the standard deviation, otherwise it becomes really hard to solve) and you want to somehow end up with something looking like this (**two-tailed**):

TWO-TAILED PROBABILITIES (practically appears everywhere):

$$P\left(\left|\frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}}\right| < c\right)$$

In other words, it is $P(h(|y - \mu|) < h(\widehat{y} - \mu))$ where $h(x) = x / \left|\frac{s}{\sqrt{n}}\right|$.

Or this (no absolute value, hence **one-tailed**):

ONE-TAILED PROBABILITY (typically does not appear in practice)

$$P\left(\frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} < c\right)$$

Where c is any number you prefer (within reason).

Note: the probability values from the T-table are TAILS, meaning you need to have it be subtracted from 1 to get the probability you actually want.

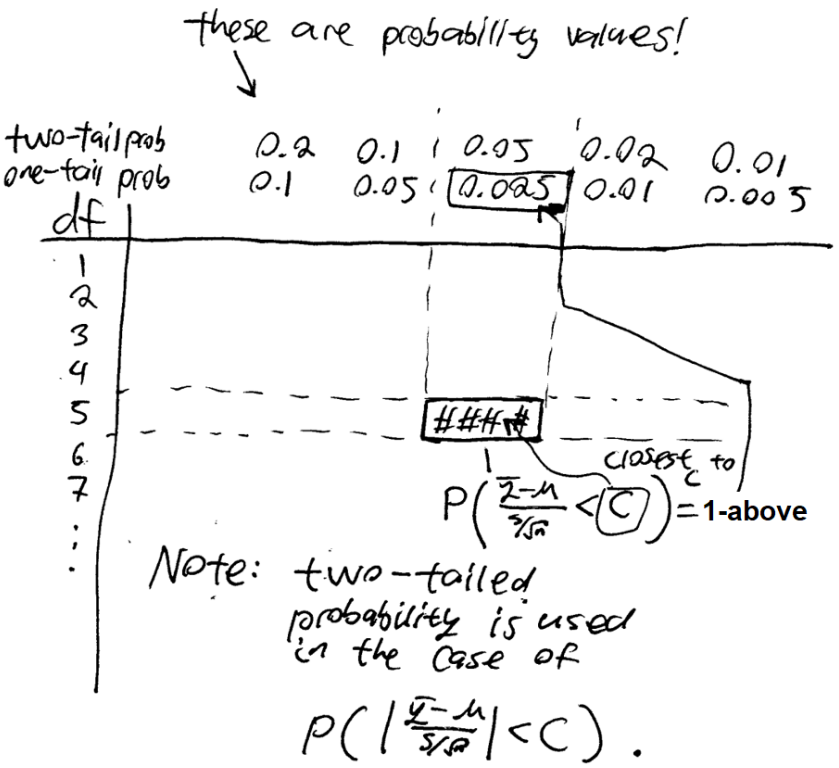

Figure 4: How to use the T-distribution table

22.1 Example

We chose 8 random samples from Y, and we call our selected samples $Y_1, Y_2, \ldots, Y_8$. We can estimate the population mean and variance:

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n} \quad \mu_{\bar{y}} = \mu$$

This means that $\sigma_{\bar{y}}^2$ can be estimated by $\frac{s^2}{n}$. Hence, if we want to find the probability that $\bar{y}$ will be within 2 standard deviations $2 \cdot \frac{s}{\sqrt{n}}$ of the theoretical population mean μ, we'll have to start with:

$$P\left(|\overline{y}-\mu| \leq 2 \cdot \frac{s}{\sqrt{n}}\right)$$
$$= P\left(\frac{|\overline{y}-\mu|}{\frac{s}{\sqrt{n}}} \leq 2\right)$$

23 Chi-Squared

The chi-squared distribution is used to find confidence intervals **of the possible sample variance** given theoretical variance and sample size. **Given a probability (the "confidence" of the interval you want to construct), you'll need to find the interval itself.**

If I have n random variables, each following the normal distribution:

$$y_1, y_2, \ldots, y_n \overset{\text{all following}}{\sim} N(\mu, \sigma)$$

Then, I can say that:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

PATTERN MATCH!! The degrees of freedom is always $\underset{\text{sample size}}{n} - 1$.

A typical question will look like: find b_1 and b_2 such that $P\left(b_1 \leq s^2 \leq b_2\right)$. Mold this such that the middle term looks like $\frac{(n-1)s^2}{\sigma^2}$, then you can replace it with χ^2_{n-1}. You should end up with something looking like this:

$$P\left(h\left(b_1\right) \leq \chi^2_{n-1} \leq h\left(b_2\right)\right)$$

Where:

$$h\left(s^2\right) = \frac{(n-1)s^2}{\sigma^2}$$
$$\Rightarrow h(b) = \frac{(n-1)b}{\sigma^2}$$

Let $t,\ u$ be any $\mathbb{R}^{\geq 0}$ such that $t+u = 1 - P\left(b_1 \leq s^2 \leq b_2\right)$. (likely 0.90, or depending on the context, if the person you wish to satisfy asks for 90% confidence intervals or 95% confidence intervals)

For $h\left(b_1\right)$, let it equal to, in the table, row first column second, $\left(\text{degrees of freedom},\ \chi^2_{1-t}\right)$. For $h\left(b_2\right)$, let it be equal to $\left(\text{degrees of freedom},\ \chi^2_{u}\right)$. You should end up with:

$$h\left(b_1\right) = r$$
$$h\left(b_2\right) = s$$

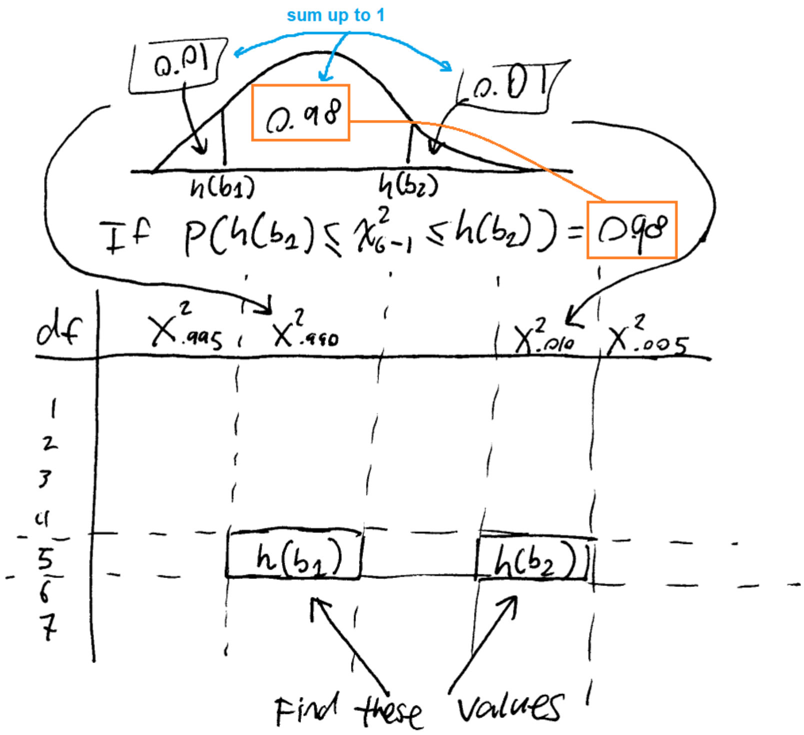

Figure 5: How to use the chi-squared table. This example has a sample size of 6, so don't confuse it with the standard deviation.

Then

$$b_1 = h^{-1}(r)$$
$$b_2 = h^{-1}(s)$$

The lower and upper bounds of the interval are the sample mean plus and minus the critical values (r, s) divided by the square root of the sample size (basically h^{-1}), respectively.

24 F-Distribution

We can also call this the F-ratio, which relates the variances of independent samples. Lucky if we get two samples with the exact same theoretical variances, then we can work with these types of questions very easily. **Key word: ratio between two sample variances.**

We let:

- W_1 be a χ^2-distributed random variable with v_1 degrees of freedom
- W_2 be another χ^2-distributed random variable with v_2 degrees of freedom, independent of W_1 completely

All we're asking is that such random variable is χ^2 distributed.

Then, we say that F, represented as this:

$$F = \frac{\frac{W_1}{v_1}}{\frac{W_2}{v_2}}$$

Is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom. Okay, that's a bit to take in:

$$F_{v_2}^{v_1}$$

That's how we denote this distribution exactly and unambiguously. Alternatively, we say that this distribution has (v_1, v_2) degrees of freedom associated with it.

A question surrounding the F-distribution will likely look like this:

$$P\left(F_{v_2}^{v_1} \leq b\right) = 0.95, \text{ solve for } b$$

This question asks for the critical values of F at the $p = 0.05$ level of significance, as $1 - 0.95 = 0.05$ (hence, 95% chance that something didn't occur by random chance alone). Hence, you need to use the correct version of the F table.

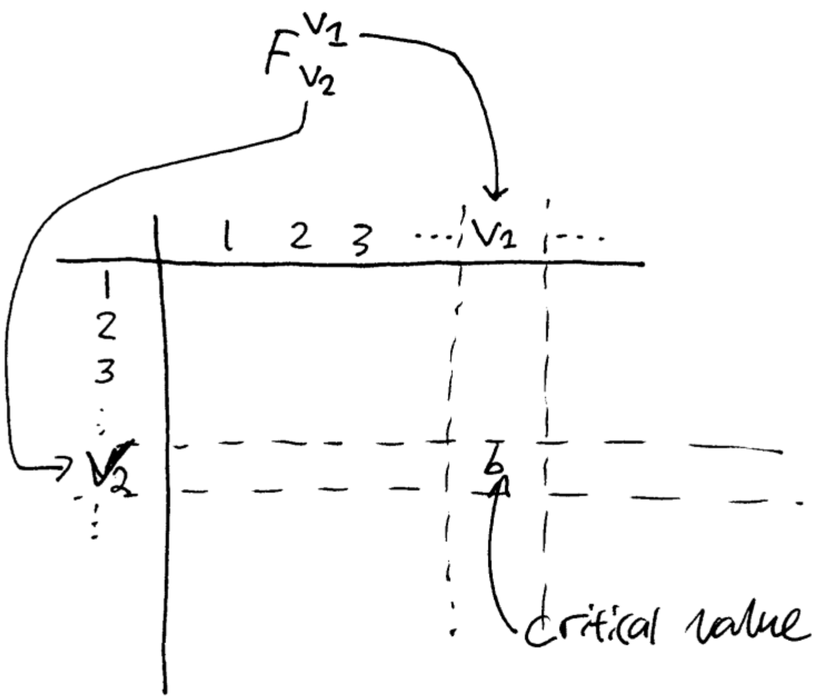

Figure 6: Using the F-table

25 Chebyshev's Inequality

Chebyshev's inequality guarantees that no more than $\frac{1}{k^2}$ of a distribution's values can be k or more standard deviations away from the mean.

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} V(X)$$

This is messy, so let $\varepsilon = k\sigma$ (we can say k standard deviations). Then:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

We can swap the direction of the inequality:

$$P(|X-\mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

Interpretation. $|X-\mu|$ is X's distance from the mean. The probability that it is less than k standard deviations away from the mean is always greater than $1-\frac{1}{k^2}$. This is a lower bound – a quick estimator!

26 The Law of Large Numbers

Tending towards the mean.

26.1 The Weak Law

Let $X_1, X_2, \ldots$ be an independent sequence of random variables with finite mean μ and variance σ^2. $\forall n \in \mathbb{N}^{\geq 1}$, let $S_n = X_1 + \cdots + X_n$. Then, $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0$$

It is trying to say:

$$\overline{X}_n \overset{\text{in probability}}{\to} \mu \quad \text{when } n \to \infty$$

26.2 The Strong Law

Let $X_1, X_2, \ldots$ be an independent sequence of random variables with finite mean μ. Then:

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1$$

It's trying to say that $\frac{S_n}{n}$ converges to μ with probablty 1 as n approaches infinity – that, almost surely.

$$\overline{X}_n \overset{\text{almost surely}}{\rightarrow} \mu \quad \text{when } n \to \infty$$

26.3 Debunking The Difference

They're trying to say the same thing:

- The weak law describes the probability of the distance from the mean being greater than any ε, which is 0, if n is big enough.
- The strong law describes the probability of the sample mean being the theoretical mean as n approaches infinity: 1, or "almost surely" (I'm not saying guaranteed).